Article

A human gut bacterial genome and culture collection for improved metagenomic analyses

Forster, Samuel C., Kumar, Nitin, Anonye, Blessing O., Almeida, Alexandre, Viciani, Elisa, Stares, Mark D., Dunn, Matthew, Mkandawire, Tapoka T., Zhu, Ana, Shao, Yan, Pike, Lindsay J., Louie, Thomas, Browne, Hilary P., Mitchell, Alex L., Neville, B. Anne, Finn, Robert D. and Lawley, Trevor D.

Forster, Samuel C., Kumar, Nitin, Anonye, Blessing O., Almeida, Alexandre, Viciani, Elisa, Stares, Mark D., Dunn, Matthew, Mkandawire, Tapoka T., Zhu, Ana et al (2019) A human gut bacterial genome and culture collection for improved metagenomic analyses. Nature Biotechnology, 37 . pp. 186-192. ISSN 1087-0156

RESOURCE
https://doi.org/10.1038/s41587-018-0009-7

nature biotechnology

OPEN

A human gut bacterial genome and culture collection for improved metagenomic analyses

Samuel C. Forster[1,2,3,8*], Nitin Kumar[1,8], Blessing O. Anonye[1,7], Alexandre Almeida[4,5], Elisa Viciani[1], Mark D. Stares[1], Matthew Dunn[1], Tapoka T. Mkandawire[1], Ana Zhu[1], Yan Shao[1], Lindsay J. Pike[1], Thomas Louie[6], Hilary P. Browne[1], Alex L. Mitchell[4], B. Anne Neville[1], Robert D. Finn[4] and Trevor D. Lawley[1*]

Understanding gut microbiome functions requires cultivated bacteria for experimental validation and reference bacterial genome sequences to interpret metagenome datasets and guide functional analyses. We present the Human Gastrointestinal Bacteria Culture Collection (HBC), a comprehensive set of 737 whole-genome-sequenced bacterial isolates, representing 273 species (105 novel species) from 31 families found in the human gastrointestinal microbiota. The HBC increases the number of bacterial genomes derived from human gastrointestinal microbiota by 37%. The resulting global Human Gastrointestinal Bacteria Genome Collection (HGG) classifies 83% of genera by abundance across 13,490 shotgun-sequenced metagenomic samples, improves taxonomic classification by 61% compared to the Human Microbiome Project (HMP) genome collection and achieves subspecies-level classification for almost 50% of sequences. The improved resource of gastrointestinal bacterial reference sequences circumvents dependence on de novo assembly of metagenomes and enables accurate and cost-effective shotgun metagenomic analyses of human gastrointestinal microbiota.

The human gastrointestinal tract harbors a diverse and dynamic microbial community that directly impacts human biology and health[1–3]. This complex ecosystem is dominated by bacteria, but also includes viruses, archaea, fungi and other eukaryotes. Metagenomic sequencing is the main method used to study GI tract microbiomes and other microbiomes in both natural and built environments[1,2,4]. Amplicon sequencing, targeting the 16S ribosomal RNA (rRNA) gene, enables characterization of taxonomic level bacterial and archaeal compositions and can detect structural changes in microbial communities. However, biologically relevant phenotypic differences exist even between highly related bacterial strains of the same species[5], and these strain differences cannot typically be distinguished by amplicon sequencing. Using shotgun metagenomic sequencing, it is feasible to assess the entire genomic content of any microbiome and achieve precise taxonomic classification and accurate functional assignments, but only if metagenome sequences can be interpreted to reveal all the species and strains present[1].

Computational approaches can be applied to extract species- and even subspecies-level information from metagenomic samples[1–3,6–10]; however, these approaches are fundamentally constrained by their requirement for deep sequence coverage and the inability to differentiate between closely related bacterial taxa[11]. Furthermore, genomes derived from de novo metagenomic assemblies may be incomplete or may represent chimeric species populations, unlike high-quality reference genomes generated from pure cultures[12]. These factors limit the accuracy of high-resolution taxonomic classification and functional analysis using metagenome-derived genomes. There is evidence that many people harbor multiple strains of the same bacterial species in their gastrointestinal microbiota[6], which means there is a pressing need to improve the precision and accuracy of metagenomic analyses to enable the functional validation that is required to develop microbiome-based therapeutics[13,14].

Comprehensive collections of reference-quality bacterial genomes enable accurate, reference-based metagenomic analysis (RBMA) and achieve species-, subspecies- and strain-level taxonomic classification of the bacterial composition of a microbiome. Substantial effort has been devoted to assembling bacterial reference genomes from different environments[15] including the Human Microbiome Project (HMP), which has sequenced bacterial isolates from 18 human body sites[16]; however, due to the diversity between individuals and previous limits in culturing methods, the majority of species still remain to be isolated, archived and genome sequenced. With recent advances in bacterial culturing methods, it is now possible to grow and purify most bacteria from the human GI tract in the laboratory[17–20].

In addition to genome sequences, access to archived bacterial isolates for functional experiments facilitates the transition from sequence-based, correlative studies to causative phenotypic validation of predicted bacterial function[13]. We report compilation and sequencing of the Human Gastrointestinal Bacteria Culture Collection (HBC), which contains isolates from the human GI tract and should enable accurate metagenomic analyses without a requirement for de novo assembly or ultra-deep sequencing and experimental validation.

[1]Host-Microbiota Interactions Laboratory, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [2]Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research, Clayton, Victoria, Australia. [3]Department of Molecular and Translational Sciences, Monash University, Clayton, Victoria, Australia. [4]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. [5]Bacterial Genomics and Evolution Laboratory, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [6]Department of Microbiology and Infectious Diseases, University of Calgary, Calgary, Alberta, Canada. [7]Present address: Microbiology and Infection Unit, Division of Biomedical Sciences, Warwick Medical School, University of Warwick, Coventry, UK. [8]These authors contributed equally: Samuel C. Forster, Nitin Kumar.
*e-mail: sf15@sanger.ac.uk; tl2@sanger.ac.uk

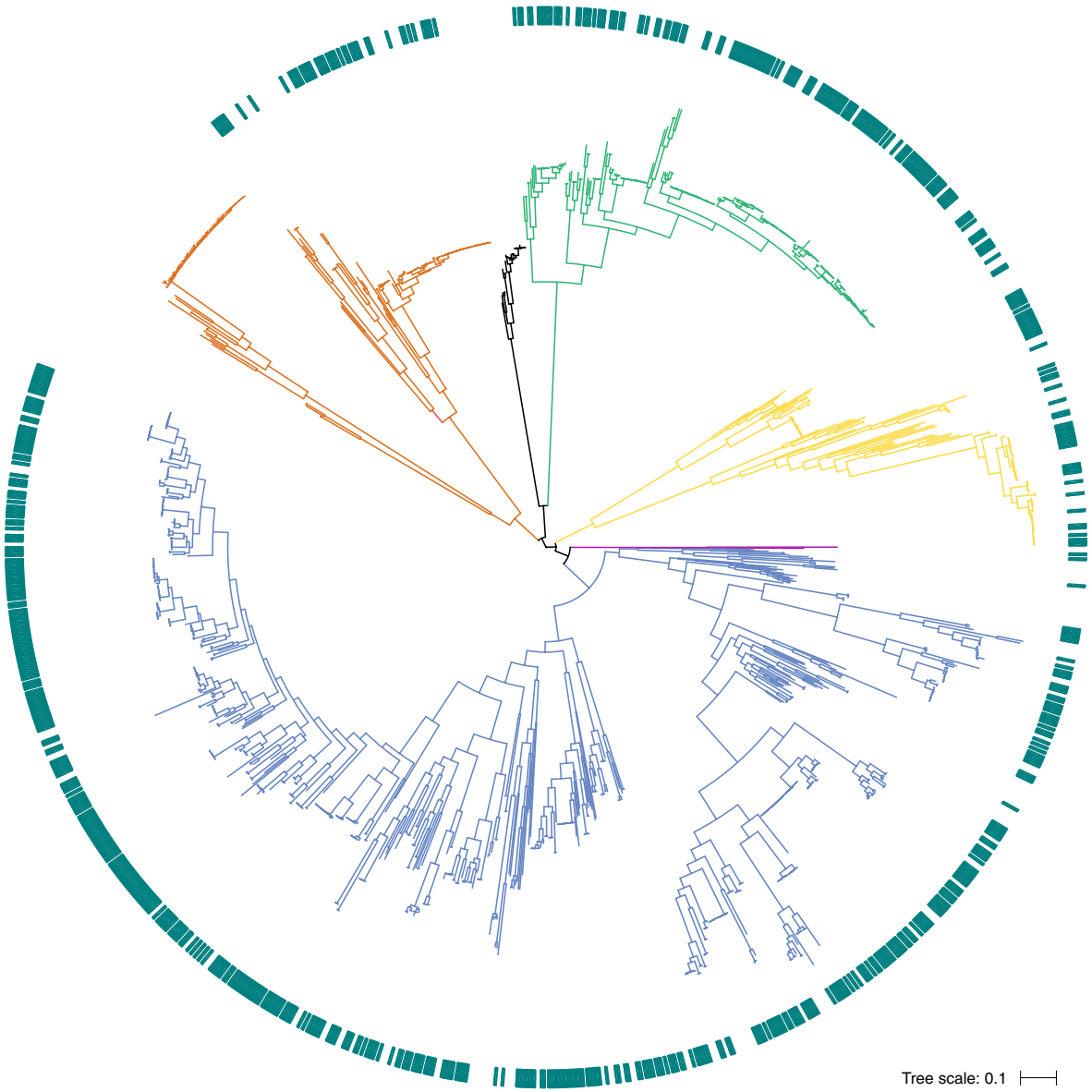

Fig. 1 | Phylogenetic diversity of the human gastrointestinal microbiota genome collection. Maximum-likelihood tree generated using the 40 universal core genes from the 737 HBC genomes (green outer circle) and the 617 high-quality public genomes derived from human gastrointestinal tract samples, which together make up the HGG. Branch color distinguishes bacterial phyla belonging to Actinobacteria (gold; $n=129$ genomes), Bacteroidetes (green; $n=231$ genomes), Firmicutes (blue; $n=772$ genomes), Fusobacteria (black; $n=26$ genomes), Synergistetes (pink; $n=2$ genomes) and Proteobacteria (orange; $n=194$ genomes) shown.

Results

Assembly of a gastrointestinal bacteria culture collection. To assemble a comprehensive collection of bacterial isolates from the human GI tract, we cultured and purified bacterial isolates from fecal samples of 20 adults based in the United Kingdom ($n=8$) and North America ($n=12$). In total, we picked more than 10,000 bacterial isolates that were then taxonomically classified using 16 S rRNA gene sequencing. Combined with 234 GI tract isolates that we previously reported[17], 737 purified and archived isolates are now included in the HBC. This collection represents 273 species (105 novel species) from 31 families in the phyla Actinobacteria (53 genomes; 16 species), Bacteroidetes (143 genomes; 40 species), Firmicutes (496 genomes; 203 species) and Proteobacteria (45 genomes; 14 species) (Supplementary Table 1). A genome sequence is available for each isolate in the HBC.

We combined our HBC genomes with 617 publicly available, high-quality human gastrointestinal-associated bacterial genomes available through the National Center for Biotechnology Information (NCBI) genome database to generate the Human Gastrointestinal Microbiota Genome Collection (HGG; Supplementary Table 1). Notably, 53% of species represented in the HGG genomes are archived in the HBC. Many of the remaining species currently absent from the HBC but present in the HGG include members of the Fusobacteria, Proteobacteria and Synergistetes, which are typically absent from healthy individuals in the developed world. This suggests further targeted culturing is required from a more diverse cohort of healthy donors and those affected with disease to exhaustively archive the bacterial component of the human GI tract microbiota.

In total, the 1,354 genomes in the HGG represent 530 species from 57 families within the phyla Actinobacteria (129 genomes;

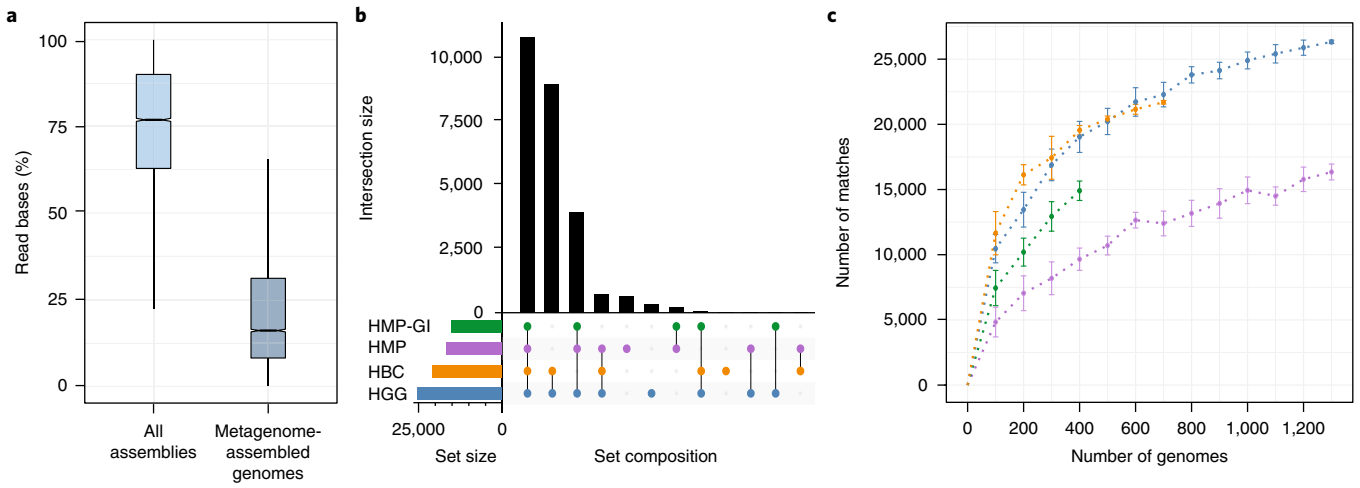

Fig. 2 | Comparison of high-quality reference genomes from de novo assembly and HGG. a, Read base usage as a percentage of total read bases present within the metagenomics samples ($n = 13,490$) that could be mapped to their respective de novo assembled contigs (min., 22.23; Q1: 62.87; median, 76.89; Q3, 89.99; max., 99.98) and metagenome-assembled genomes (MAGs; min., 0.16; Q1, 8.17; median, 16.09; Q3, 31.16; max., 65.64). **b**, Total number of classified bins using HGG (blue), genomes derived from the HBC collection alone (HBC; orange), the HMP (purple) and gastrointestinal derived isolates from the HMP (HMP-GI; green). **c**, Total number of 39,913 MAGs classified using subsampled sets of genomes from the HGG (blue), HBC (orange), HMP (purple) and HMP-GI (green). Error bars show mean and s.d. ($n = 100$ bootstraps).

55 species), Bacteroidetes (231 genomes; 69 species), Firmicutes (772 genomes; 339 species), Fusobacteria (26 genomes; 9 species), Proteobacteria (194 genomes; 56 species) and Synergistetes (2 genomes; 2 species) (Supplementary Fig. 1). To understand the phylogenetic relationship between these taxa, we extracted 40 universal core genes[21] from each genome and performed phylogenetic analysis (Fig. 1; Supplementary Fig. 2). Overall, the maximum phylogenetic diversity was observed in Firmicutes, particularly the classes Clostridia, Erysipelotrichia and Negativicutes; however, a broad range of species and phylogenetic group are represented across all phyla (Fig. 1; Supplementary Fig. 2).

HGG improves gastrointestinal metagenome analyses. In the absence of reference genomes, state-of-the-art analysis of metagenomic sequencing is dependent on de novo assembly of raw reads, followed by contig binning to generate metagenome-assembled genome sequences (MAGs). To compare the efficiency of taxonomic classification of de novo assembly and binning to RBMA analysis, we considered 13,490 publicly accessible (Supplementary Table 2) shotgun metagenomes from feces, with sufficient read coverage to perform de novo assembly. De novo assembly and contig binning identified 11,892 samples (88.2%) that were of sufficient quality to produce contigs greater than or equal to 2,000 base pairs (bp) in length. A total of 39,913 bins with >90% completeness and <5% contamination, referred to hereafter as MAGs, were obtained from 9,548 assemblies (Supplementary Table 3). Of these MAGs, 81% had at least 15 tRNAs, further emphasizing their high level of completeness; however, only 16.1% (interquartile range, IQR = 31.2%–8.2%) of read bases contributed to these MAGs (Fig. 2a).

To evaluate how the HBC genome collection alone and the complete HGG collection compares to the existing HMP genomes, we next considered which of the MAGs could be identified using each collection as a reference database. The HGG was able to identify 25,085 MAGs compared to 20,772 with the genomes corresponding to just the HBC. Meanwhile, 16,476 MAGs were identified with the HMP collection from all 18 body sites, and 15,156 MAGs were found when including only HMP isolates from the gastrointestinal body sites (HMP-GI). This represents a 52.3% improvement when using the HGG collection as a reference compared to the complete HMP (Fig. 2b). As the HGG collection is considerably larger than the HBC, HMP and HMP-GI genome collections, we next performed bootstrapped subsampling of each genome database and compared the selected genomes to the previously identified MAGs by average nucleotide identity (ANI > 95%). Considering subsamples of 400 genomes, the last data point available for the HMP-GI, the HGG achieves 19,545 matches and the HBC genome collection 19,036 matches compared to 14,906 matches with the HMP-GI and 9,655 with the full HMP (Fig. 2c). Classification is hindered in the full HMP, as it includes genomes from nongastrointestinal species. Notably, the greater matching achieved using the HGG and HBC genomes suggests more representative phylogenetic diversity is also present within these datasets. Thus, our analysis demonstrates a 61.1% increase in classification potential with the HGG compared to the existing genomes.

Phylogeny-based estimate of genome coverage in metagenomes. Although it is possible to generate MAGs using de novo assembly and binning approaches, this method remains unable to assign 83.9% of reads considered within the 13,490 shotgun metagenomic sequenced samples analyzed in this study. To address this limitation, we next compared all de novo assembled contigs with the HGG to determine the ability to classify a larger proportion of the input data. Applying this method, we were able to map 74.5% (IQR = 84.1%–62.9%) of contigs at a level approximately equivalent to genus (90% cutoff), whereas we could assign 67.3% (IQR = 78.7%–54.8%) at the species level (95% cutoff; Fig. 3a). Remarkably, 40.8% (54.3% – 30.0%) could be classified below species level (99% cutoff) despite not including any isolates cultured from any of these samples in the HGG (Fig. 3a).

Given the improvements in classification that the HGG provides, we next adopted a lowest common ancestor RBMA to determine overall taxonomic classification efficiency across the same dataset. Compared to de novo metagenomic assembly and binning approaches, RBMA is more resilient to low sample coverage because it requires shallower sequencing depth to confidently assign a sequence to a reference genome. With these datasets, RBMA of large-scale shotgun metagenomic datasets required a median processing time of 7.3 min for each sample compared to the 12.19 h required for an equivalent de novo assembly. This substantial reduction in required computational performance provides a means to

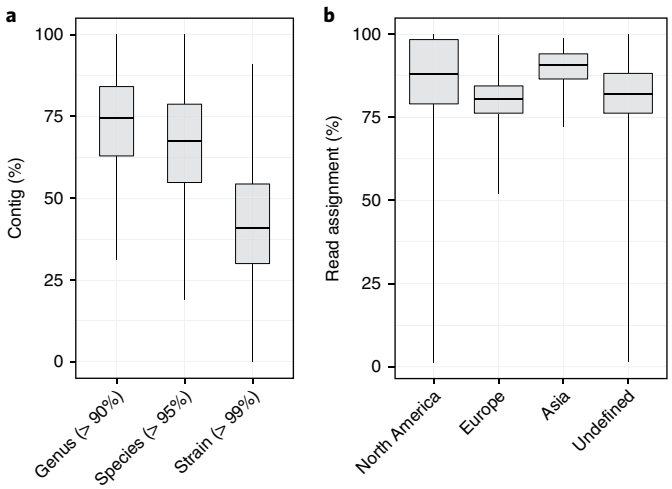

Fig. 3 | Classification efficiency using the HGG. a, Contig assignment from 13,490 metagenomic samples at genus (90%; min., 31.35; Q1, 62.92; median, 74.48; Q3, 84.10; max., 100.0), species (95%; min., 18.94; Q1, 54.80; median, 67.35; Q3, 78.73; max., 100.0) and strain (99%; min., 0.0; Q1, 30.03; median, 40.82; Q3, 54.35; max., 90.77) identity compared to the HGG. **b**, Classification of metagenomic sequenced samples from North America ($n=2,064$; min., 1.31; Q1, 79.07; median, 88.16; Q3, 98.42; max., 99.97), Europe ($n=1,431$; min., 52.07; Q1, 76.28; median, 80.66; Q3, 84.47; max., 99.52), Asia ($n=191$; min., 72.37; Q1, 86.56; median, 90.84; Q3, 94.13; max., 98.93) and the other undefined locations ($n=9,804$; min., 1.45; Q1, 76.28; median, 82.14; Q3, 88.25; max., 99.94).

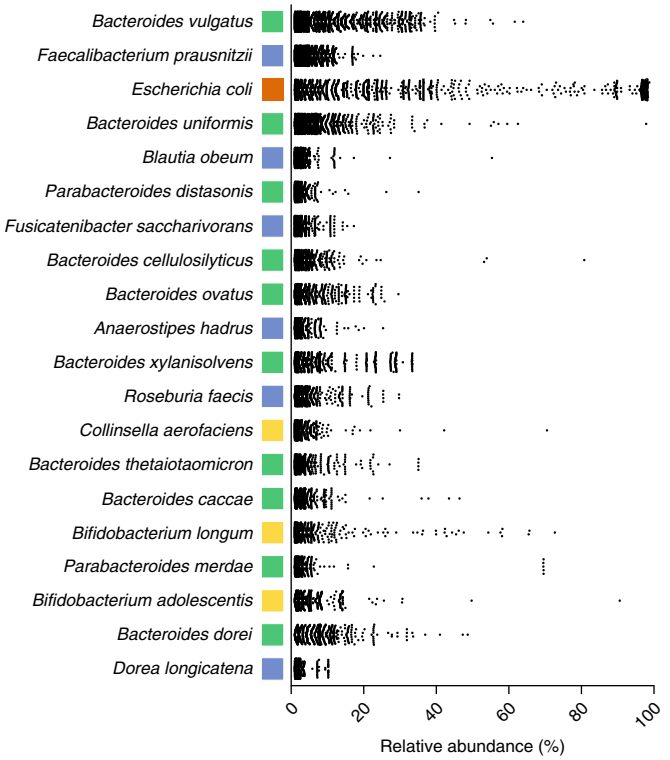

Fig. 4 | Dominant bacterial species within the human gastrointestinal microbiota. Dominant species, ordered by prevalence, found within the 13,490 human gastrointestinal metagenomic samples and their relative abundance within each sample. Color denotes Bacteroidetes (green), Firmicutes (blue), Proteobacteria (orange), Actinobacteria (gold).

process more samples and overcome the limitations in statistical power that hinders many metagenomic studies.

Horizontal gene transfer of mobile elements in bacterial populations and communities can limit our ability to identify the true species composition because of incorrect read assignment. To address confounding factors associated with horizontal gene transfer and provide a more precise estimate of taxonomic coverage, we also generated a comprehensive list of mobile elements, insertion sequences and plasmids found within the European Nucleotide Archive (ENA)[22]. Combined with the mobile elements predicted in the HGG, this represents a comprehensive database of known mobile elements found within the human gastrointestinal microbiota. The database includes 2,803 plasmids and 489 transposons and insertion sequences that were masked within the genomes and filtered from metagenomic reads before lowest common ancestor classification to maximize the phylogenetic signal (Supplementary Table 4). When we applied the lowest common ancestor RBMA with the mobile element filtered HGG, classification of the raw reads achieved an average taxonomic assignment of 82.9% at the genus level and 78.7% at species levels. Taken together, these analyses reveal that high-resolution classification of the majority of metagenomic reads derived from the human gastrointestinal microbiota can be achieved using the HGG even when considering samples across diverse geographic populations (Fig. 3b).

Bacterial diversity in the human gastrointestinal tract. We next sought to understand which species were most prevalent within the human gastrointestinal microbiota using the HGG. To do this, we interrogated all of the 13,940 high-quality shotgun metagenomic samples derived from human feces (Supplementary Table 2). Though this analysis may be impacted by variation in fecal sample storage conditions and DNA extraction methods[23], we reasoned that those species that are highly prevalent across samples from many individuals are likely to play an important role in human biology and should be the focus of further investigation. Considering only species that are present at a level greater than 0.01% within any sample, we identified 165 species present across more than two unrelated samples (Supplementary Table 5). This group of dominant species included Bacteroidetes ($n=41$), Firmicutes ($n=82$), Proteobacteria ($n=27$) and Actinobacteria ($n=15$). Given the background prevalence of each phylum, this represents a significant overrepresentation of species from Bacteroidetes ($P<0.05$) and a significant underrepresentation of species from Firmicutes ($P<0.01$).

Considering all species that were detected above background levels, the majority of dominant species remain as members of the Bacteroidetes. In total, 8 of the top 20 prevalent species were members of the *Bacteroides* genus (*Bacteroides vulgatus*, *Bacteroides uniformis*, *Bacteroides cellulosilyticus*, *Bacteroides ovatus*, *Bacteroides xylanisolvens*, *Bacteroides thetaiotaomicron*, *Bacteroides caccae* and *Bacteroides dorei*). When corrected for the number of species within each phylogenetic group, the Bacteroidetes generally, and the *Bacteroides* and *Parabacteroides* genera (*Parabacteroides distasonis* and *Parabacteroides merdae*) more specifically, are significantly overrepresented ($P<0.001$; Fig. 4). Despite three being over 346 species within the Firmicutes, there are only 6 distantly related Firmicute species that were highly represented across many individuals (*Fecalibacterium prausnitzii*, *Blautia obeum*, *Fusicatenibacter saccharivorans*, *Anaerostipes hadrus*, *Roseburia faecis* and *Dorea longicatena*; Fig. 4). Overall, all detected genera within the Firmicutes phylum were statistically underrepresented in their occurrence. Similarly, the only member of the Proteobacteria that was highly prevalent across samples was *Escherichia coli*, with the majority of Proteobacteria not detected within the samples. Interestingly, no members of the Fusobacteria or Synergistetes were found to be

prevalent at the level of detection considered, suggesting that they are found only during certain conditions or stages of life that were not included in this analysis.

These data suggest a potential key role for specific members of the Bacteroides within the human gastrointestinal tract. In contrast, the significantly greater diversity observed in Firmicutes, the other dominant phyla, suggests a highly variable, potentially functionally redundant group consistent with previous reports of dynamic spore-mediated transmission and turnover for many taxa within this group[17,24]. Although laboratory-based phenotypic analysis examining many of the key species that were identified through this study remains limited, this can now be addressed through access to the isolates archived in the HBC.

Given the diverse array of novel genomes contained within the HGG, we next sought to understand the prevalence of these species across the community. Importantly, the availability of these genomes allows us to reliably assess the prevalence of these species in metagenome samples for the first time. In total, 106 of the 173 novel genomes (60.9%) are found at greater than 0.001% abundance in at least one sample within the 13,490 public metagenome samples available. Notably, almost half (87; 48.6%) were found in >100 samples, but less than one quarter (39; 21.8%) were found in >1,000 samples. Interestingly three novel species all within the Clostridiales were found in almost half the samples analyzed. Two novel Lachnospiraceae were respectively found in 7,797 (55.9%) and 7,074 (50.7%) samples, and a new Ruminococcaceae species was found in 6,777 (48.6%) samples. Collectively, these data suggest many of the novel species and genomes identified through this work occur frequently within the human population and potentially represent integral parts of the human gastrointestinal microbiota that warrant further investigation.

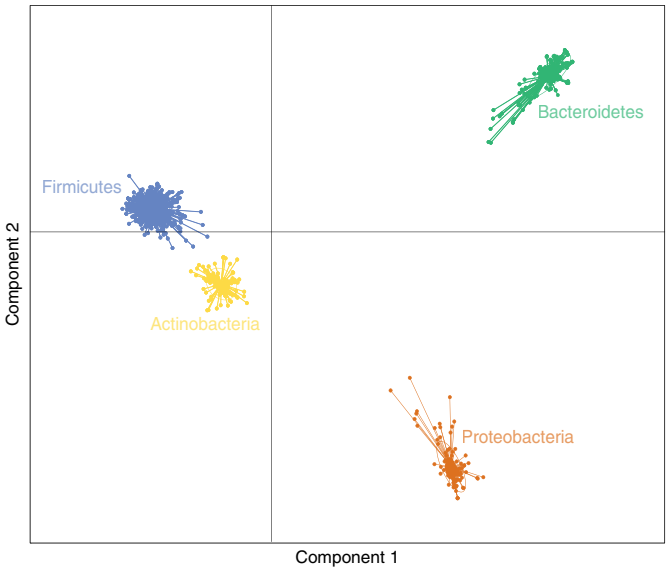

Fig. 5 | Bacterial functions in the human gastrointestinal tract. DAPC analysis of functional categories shows a clear separation of functions associated with each dominant phylum (Bacteroidetes (green; $n=231$ genomes), Firmicutes (blue; 772 genomes), Proteobacteria (orange; $n=194$ genomes), Actinobacteria (gold; $n=129$ genomes)) within the HGG collection.

Functions of human gastrointestinal bacteria. This extended collection of genome sequenced bacterial isolates enables high-resolution functional and taxonomic analysis. We first performed a clusters of orthologous group of protein (COG) annotation[25] on the protein sequences to identify those features prevalent within the HGG bacteria. This analysis identified 4,696 distinct orthologous groups represented in at least one isolate. As expected, bacterial housekeeping functions, including ribosomal protein function, amino acid synthesis and other translation-associated functions, dominate the 30 functions found in all bacteria within the collection (Supplementary Table 6).

To understand differences in the functional role performed by members of the four major bacterial phyla of the gastrointestinal microbiota (Bacteroidetes, Firmicutes, Actinobacteria and Proteobacteria), we compared 4,696 orthologous groups identified with COG analysis using discriminant analysis of principle components (DAPC). This comparison demonstrates clear functional differences between key phyla of the human gastrointestinal microbiota (Fig. 5). Next, we undertook an enrichment analysis to identify those functions overrepresented in each phylum relative to all functions present within the HGG. This analysis identified 8, 122, 152 and 389 statistically enriched functions ($q<0.001$) in Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria, respectively (Supplementary Table 7). Enriched functions within the Actinobacteria were limited, but those identified were primarily associated with lipid ($q<1.99\times10^{-83}$) and carbohydrate metabolism ($q<7.57\times10^{-77}$). Equivalent analysis of the Bacteroidetes specific functions identified many key functions, including iron ($q<1.18\times10^{-114}$) and sulfur transporter functions ($q<6.82\times10^{-97}$) and specific sodium-transporting NADH ubiquinone oxidoreductases ($q<3.47\times10^{-124}$). Firmicutes were dominated by uncharacterized functions; however, spore formation ($q<3.48\times10^{-123}$) and thiamine ($q<2.76\times10^{-101}$) and riboflavin ($q<7.04\times10^{-101}$) transport were all highly enriched. Finally, Proteobacteria were dominated by fructose bisphosphatase ($q<4.50\times10^{-140}$), glucokinases ($q<4.55\times10^{-125}$) and regulators of iron cluster formation ($q<9.20\times10^{-98}$). These results demonstrate the distinct differences in the unique functions provided by the key phyla of the human gastrointestinal microbiota; however, the prevalence of uncharacterized functions further demonstrates the need for better genome annotation and functional genomics to understand these bacteria.

The HGG collection contains genomes from 173 species not previously isolated from the human gastrointestinal tract. This includes genomes from the 105 novel species within the HBC and genomes from 68 known species in which genome-sequenced isolates from the human gastrointestinal tract did not previously exist (Supplementary Table 1). To understand what functions were found within these 173 species but were absent within the previously reported genome-sequenced species, we performed functional analysis. In total, 45 newly described functions, of which 41 were found in the Firmicutes, were identified. Though these functions were dominated by uncharacterized proteins, novel functions included those associated with tetrahydromethanopterin *S*-methyltransferase (present in five species), preprotein translocase (also present in five species) and formaldehyde-activating enzyme necessary for methanogenesis (found in four of the previously uncharacterized Firmicutes). In addition, 83.2% of these newly sequenced isolates and 85.8% of the novel species are predicted to form spores on the basis of previously defined genomic signatures[17].

Finally, we sought to understand which functions were predicted to occur in newly genome-sequenced members of a particular phyla but were absent in all existing genomes of that phyla. This analysis identified type III, IV and VI secretion system components in Bacteroidetes that were not found in any of the previously sequenced gastrointestinal Bacteroidetes but were recognized within the existing Proteobacteria and Firmicutes genomes. Equally, ABC transporter functions found in existing genomes from Proteobacteria were identified within the newly sequenced gastrointestinal Firmicutes but not within any of the previously sequenced isolates. This suggests further functional overlap that

may exist between specific members of phyla with potentially important, redundant roles in microbial community dynamics and host–microbiota interactions.

Discussion

We present a gastrointestinal bacteria genome and culture collection that substantially increases the proportion of species found in metagenomics samples from the developed world. The YCFA medium used achieves bacterial growth at levels broadly representative of the original sample, so it will therefore be necessary to combine YCFA with selective culturing techniques to target specific bacterial phenotypes; for example, antibiotic resistance, sporulation, carbohydrate utilization and isolation of rare bacterial species present in individual fecal samples. Shotgun metagenomic sequencing has not been performed for many of the world's population, so it is not currently possible to accurately assess the proportion of cultured bacteria across the entirety of the human population. We proposed that an expanded, coordinated, global culturing exercise with particular focus on samples and bacterial isolates from the developing world and more diverse communities across the developed world is needed. Collection and storage of metadata associated with these metagenomic samples is also essential: despite efforts to develop standards[26,27], many genome and metagenomic sequences deposited in the public sequence collections have incorrect, inconsistent, missing or limited metadata that fundamentally limits their use.

In addition to improved species classification, access to comprehensive genome sequenced isolates fundamentally alters the methods, resolution and accuracy of functional analysis. Genome-sequenced isolates enable functional capacity to be inferred from the genetic repertoire of the reference genomes. This eliminates the need to perform ultra-deep metagenomic sequencing and ensures that complete functional pathways are contained within individual bacterium. In addition to improved accuracy, this method also has the capacity to improve sensitivity for functional analysis, allowing detection of functions that, although not prevalent, may represent fundamental differences between study cohorts.

Although extensive characterization of pathogens and model organisms has dominated the past 100 years of microbiology research, study of human-health-associated commensal bacteria has lagged behind. Culturing, genome sequencing and isolate archiving, as reported here, will underpin substantially improved microbiome-based analysis of the human gastrointestinal tract, and potentially other sites[28]. Traditional microbiology methods can continue to enable access to the bacterial isolates that are sorely needed to perform experimental characterization and validation and improving our understanding of important human-associated microbial communities.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41587-018-0009-7.

Received: 11 December 2017; Accepted: 13 December 2018;
Published online: 4 February 2019

References

1. Lloyd-Price, J. et al. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).
2. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
3. Scholz, M. et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* **13**, 435–438 (2016).
4. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
5. Kamada, N., Chen, G. Y., Inohara, N. & Núñez, G. Control of pathogens and pathobionts by the gut microbiota. *Nat. Immunol.* **14**, 685–690 (2013).
6. Li, S. S. et al. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* **352**, 586–589 (2016).
7. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
8. Nielsen, H. B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
9. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
10. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).
11. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
12. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
13. Neville, B. A., Forster, S. C. & Lawley, T. D. Commensal Koch's postulates: establishing causation in human microbiota research. *Curr. Opin. Microbiol.* **42**, 47–52 (2018).
14. Walker, A. W., Duncan, S. H., Louis, P. & Flint, H. J. Phylogeny, culturing, and metagenomics of the human gut microbiota. *Trends Microbiol.* **22**, 267–274 (2014).
15. Mukherjee, S. et al. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.* **35**, 676–683 (2017).
16. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
17. Browne, H. P. et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543–546 (2016).
18. Lagier, J. C. et al. Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat. Microbiol.* **1**, 16203 (2016).
19. Goodman, A. L. et al. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl Acad. Sci. USA* **108**, 6252–6257 (2011).
20. Lau, J. T. et al. Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. *Genome Med.* **8**, 72 (2016).
21. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of prokaryotic species. *Nat. Methods* **10**, 881–884 (2013).
22. Silvester, N. et al. The european nucleotide archive in 2017. *Nucleic Acids Res.* **46**, D36–D40 (2017).
23. Costea, P. I. et al. Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
24. Browne, H. P., Neville, B. A., Forster, S. C. & Lawley, T. D. Transmission of the gut microbiota: spreading of health. *Nat. Rev. Microbiol.* **15**, 531–543 (2017).
25. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
26. Yilmaz, P. et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **29**, 415–420 (2011).
27. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
28. Thomas-White, K. et al. Culturing of female bladder bacteria reveals an interconnected urogenital microbiota. *Nat. Commun.* **9**, 1557 (2018).

Acknowledgements

This work was supported by the Wellcome Trust (098051); the United Kingdom Medical Research Council (PF451 to T.L.), the BBSRC (BB/M011755/1 to R.D.F.), the European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI) and the Australian National Health and Medical Research Council (1141564 to S.C.F.). S.C.F. is supported by the Australian National Health and Medical Research Council (1091097) and the Victorian Government's Operational Infrastructure Support Program. L.J.P. is supported by a Medical Research Council Doctoral Training Grant (MR/J004111/1). The authors would also like to acknowledge the support of the Wellcome Sanger Institute Pathogen Informatics and Core Sequencing Teams.

Author contributions

S.C.F., B.A.N., N.K., R.D.F. and T.D.L. conceived the study. S.C.F., N.K., A.A., B.A.N., H.P.B., E.V., R.D.F. and T.D.L. and wrote the manuscript. S.C.F., B.A.N., B.O.A., E.V., M.D.S., M.D., H.P.B., Y.S., L.J.P. and T.L. collected samples, purified the bacteria and

performed genome sequencing. S.C.F., N.K., A.A., A.L.M., T.T.M., A.Z. and R.D.F. performed the computational analysis. All authors read and approved the manuscript.

Competing interests

S.C.F., B.A.N., M.D., R.D.F. and T.D.L. are either employees of or consultants to Microbiotica Pty Ltd.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-018-0009-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.C.F. or T.D.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Bacterial culturing and purification. Bacterial culturing was performed using supplemented YCFA medium[29] with or without ethanol pretreatment as described previously[17]. Briefly, sample processing and culturing took place under anaerobic conditions in a Whitley DG250 workstation at 37 °C using phosphate buffered saline and culture media incubated under anaerobic conditions for 24 h before use. Fecal samples were collected from 20 healthy adults (North America, $n=12$; United Kingdom, $n=8$) who had not taken antibiotics within the last six months. Samples were transported frozen and stored at −80 °C before culturing. Fecal samples were homogenized in reduced PBS (0.1 g stool per ml PBS), serially diluted and plated directly onto YCFA[29] agar supplemented with 0.002 g ml^{-1} each of glucose, maltose and cellobiose in (13.5-cm diameter) Petri dishes. Colonies were picked, restreaked to purity and identified using 16 S rRNA gene sequencing. Species were defined on the basis of a 16 S rRNA gene sequence identity threshold of >97.8%[30]. Isolates are available from the Wellcome Sanger Institute or the relevant public culture collection.

Genome sequencing and annotation. Genomic DNA was extracted from pelleted cells using a phenol–chloroform method described previously[31]. DNA was prepared and sequenced using the Illumina Hi-Seq platform with library fragment sizes of 200–300 bp and a read length of 100 or 125 bp at the Wellcome Sanger Institute as previously described[32]. Annotated assemblies were produced using the pipeline described previously[33]. For each sample, sequence reads were used to create multiple assemblies using Velvet v1.2 (ref. [34]) and VelvetOptimiser v2.2.5 (https://github.com/tseemann/VelvetOptimiser). An assembly improvement step was applied to the assembly with the best N50, and contigs were scaffolded using SSPACE[35] and sequence gaps filled using GapFiller[36]. Automated annotation was performed using PROKKA v1.11 (ref. [37]). Genomes with less than 400 contigs, a genome size less than 8 Mb and the presence of 16 s rRNA sequences with greater than 97.5% homology were considered pure and included in further analysis. All genomes within our collection are publicly available through the EBI European Nucleotide Archive under project accessions ERP105624 and ERP012217 (Supplementary Table 1). Public samples were included when the isolation source within the NCBI was fecal material or gastrointestinal-tract associated, and sequences were derived from pure isolates. All genomes were screened for quality as described for internal genomes, with only those passing these criteria included for further analysis.

Phylogenetic analysis. The phylogenetic analysis was conducted by extracting amino acid sequence of 40 universal core marker genes[38,39] from each genome in the bacterial collection using SpecI[21]. The protein sequences were concatenated and aligned with MAFFT v. 7.20 (ref. [40]), and maximum-likelihood trees were constructed using RAxML v. 8.2.8 (ref. [41]) with the standard LG model and 100 rapid bootstrap replicates. Trees were visualized using FastTree[42] followed by iTOL[43].

De novo metagenomic analysis. For the metagenomic analyses, we first extracted 13,490 metagenomic sequencing runs from human gut samples available in the European Nucleotide Archive (Supplementary Table 4). To evaluate the efficiency of de novo assembly approaches, raw reads were assembled using metaSPAdes v3.10.0 (ref. [44]) and subsequently binned with MetaBAT 2 (v2.12.1)[45], with a minimum contig length threshold of 2,000 bp. Sequencing coverage and read base usage was inferred by mapping the raw reads back to the assemblies or bins using BWA v0.7.16 (ref. [46]) and then retrieving the percentage of mapped read bases with SAMtools v1.5 (ref. [46]) and the jgi_summarize_bam_contig_depths function from MetaBAT 2 (ref. [45]). Ribosomal RNAs (rRNAs) were detected with INFERNAL v1.1.2 (ref. [47]) using the Rfam covariance models of the bacterial 5 S, 16 S and 23 S rRNAs. Total alignment length was inferred by the sum of all nonoverlapping hits. Each gene was considered present if more than 80% of the sequence was contained in the MAG. Transfer RNAs (tRNAs) were identified with tRNAscan-SE v2.0 (ref. [48]) using the bacterial tRNA model and default parameters. Bins with >90% completeness and < 5% contamination, estimated by CheckM[12] were further analyzed against the Human Microbiome Project (HMP), the HBC genomes and complete HGG.

Genome comparison. The complete HGG collection, the HBC-derived genomes and the HMP genomes were used for comparison to the MAGs. The complete HMP set, as well as a set of human gut-specific references retrieved from the HMP Project Catalog (https://www.hmpdacc.org/hmp/catalog/grid.php?dataset=genomic), were analyzed. For each database, Mash v2.0 (ref. [49]) was used to convert all reference genomes into a MinHash sketch (mash sketch) with default settings. Then, the Mash distance between the MAGs and each set of references was calculated to find the best match (i.e., the genome with the lowest Mash distance). Each MAG and its closest relative among the reference set were aligned with dnadiff v1.3 from MUMmer 3.23 (ref. [50]) to compare each pair in terms of percentage of aligned bases and ANI. MAGs that aligned above 75% of their total length with an ANI above 95% were considered a positive match. To further benchmark the assignment performance, we subsampled the reference genomes from each database at increments of 100 genomes and created MinHash sketches with a sketch size of 100,000 (*mash sketch -s 100,000*). We then assessed the number of MAGs that matched each subsampled set with a Mash distance below 0.05 (ANI > 95%). Data was visualized using the UpSet R package[51].

Lowest common ancestor metagenomic analysis. Lowest common ancestor analysis was performed using a custom generated Kraken database containing all genomes within the HBC. Metagenomic samples were filtered by quality using Trimmomatic 0.35 (ref. [52]) and human contaminating reads filtered by mapping to the Human reference genome (hg19) with bowtie2 (ref. [53]), with samples containing less than one million reads after filtering being discarded. Filtered sequences were classified at the genus and species levels using lowest common ancestor analysis as previously described[54].

Functional genomic analysis. To identify protein domains in a genome, we performed RPS-BLAST using COG database (accessed November 2017)[25]. All protein domains were classified in different functional categories using the COG database[25] and were used to perform discriminant analysis of principle components (DAPC)[55] implemented in the R package Adegenet v2.0.1 (ref. [56]). Domain and functional enrichment analysis was calculated using one-sided Fisher's exact test with *P* value adjusted by Hochberg method in R v. 3.2.2.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequence data is deposited in the ENA under project numbers ERP105624 and ERP012217. Bacterial isolates have been deposited at the Leibniz Institute DSMZ-German Collection of Microorganims and Cell Cultures (http://www.dsmz.de), the CCUG-Culture Collection, University of Gothenburg, Sweden (http://www.ccug.se), the Belgian Co-ordinated Collection of Microorganisms hosted by the Laboratory of Microbiology (BCCM/LMG) at Ghent University (http://bccm.belspo.be/) and at the Japan Collection of Microorganisms (JCM; http://jcm.brc.riken.jp/en/). Culture collection identifiers and ENA accession numbers for each genome are provided in Supplementary Table 1. Metagenome-assembled genomes are available from ftp://ftp.ebi.ac.uk/pub/databases/metagenomics/hgg_mags.tar.gz.

References

29. Duncan, S. H., Hold, G. L., Harmsen, H. J., Stewart, C. S. & Flint, H. J. Growth requirements and fermentation products of *Fusobacterium prausnitzii*, and a proposal to reclassify it as *Faecalibacterium prausnitzii* gen. nov., comb. nov. *Int. J. Syst. Evol. Microbiol.* **52**, 2141–2146 (2002).
30. Yarza, P. et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).
31. Green, M.R., Sambrook, J. & Sambrook, J. *Molecular Cloning: a Laboratory Manual*. 4th edn (Cold Spring Harbor Laboratory Press, 2012).
32. Harris, S. R. et al. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2010).
33. Page, A. J. et al. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb. Genom.* **2**, e000083 (2016).
34. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
35. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
36. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biol.* **13**, R56 (2012).
37. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
38. Ciccarelli, F. D. et al. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
39. Sorek, R. et al. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**, 1449–1452 (2007).
40. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
41. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
42. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
43. Letunic, I. & Bork, P. Interactive Tree Of Lifev2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475–W478 (2011).
44. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
45. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
46. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

47. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
48. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–W57 (2016).
49. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
50. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
51. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).
52. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
53. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
54. Forster, S. C. et al. HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. *Nucleic Acids Res.* **44**, D604–D609 (2016).
55. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94 (2010).
56. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).

nature research

Corresponding author(s): Samuel C Forster and Trevor D Lawley

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Illumina Sequencing data was collected using HCS 3.4.0.
Data analysis	Software used for data analysis: Velvet v1.2, VelvetOptimiser v2.2.5, SSPACE, GapFiller, PROKKA v1.11, MAFFT v. 7.20, RAxML v. 8.2.8, FastTree, iTOL, metaSPAdes v3.10.0, MetaBAT v2.12.1, BWA v0.7.16, samtools v1.5, MetaBAT 2 (jgi_summarize_bam_contig_depths), INFERNAL v1.1.2, tRNAscan-SE v2.0, CheckM, Mash v2.0, dnadiff v1.3, MUMmer v3.23, Kraken, Trimmomatic 0.35, bowtie2. R v3.2.2: R Packages: UpSet R, Adegenet v2.0.1,

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence data is deposited in the ENA under project numbers ERP105624 and ERP012217. ENA accession numbers for each genome, culture collections details and strain identifiers are provided in Supplementary Table 1. Metagenome-assembled genomes are available from ftp://ftp.ebi.ac.uk/pub/databases/metagenomics/hgg_mags.tar.gz.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed as sample size was limited by practical culturing capacity.
Data exclusions	No data was excluded from the analysis.
Replication	Isolates were cultured from all samples. Metagenomic analysis included all high-quality samples available in the European Nucleotide Archive so further replication was not possible.
Randomization	Isolates were cultured from all samples. Randomization is not relevant to this study.
Blinding	Isolates were cultured from all samples. Blinding is not relevant to this study.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	Bacterial isolates have been deposited at the Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures (http://www.dsmz.de), the CCUG-Culture Collection, University of Gothenburg, Sweden (http://www.ccug.se), the Belgian Co-ordinated Collection of Micro-organisms hosted by the Laboratory of Microbiology (BCCM/LMG) at Ghent University (http://bccm.belspo.be/) or the Japan Collection of Microorganisms (JCM; http://jcm.brc.riken.jp/en/). Additional isolates are available upon request

Human research participants

Policy information about studies involving human research participants

Population characteristics	Samples were collected from healthy individuals between the ages of 25 and 55 with no reported antibiotic administration within the last 6 months. All samples were provided anonymously.
Recruitment	Healthy donors nominated to provide faecal samples for culturing in Canada and the UK. Donors with gastrointestinal disorders, chronic health conditions or with reported antibiotic administration within the last 6 months were excluded. Sampling of more diverse communities is likely to further increase the diversity of bacteria recovered.